

Why Does Zipf's Law Break Down in Rank-Size Distribution of Cities?

Hiroto KUNINAKA * and Mitsugu MATSUSHITA

Department of Physics, Chuo University, Kasuga, Bunkyo-ku, Tokyo 112-8551

We study rank-size distribution of cities in Japan on the basis of data analysis and computer simulation. From the census data after World War II, we find that the rank-size distribution of cities is composed of two parts, each of which has independent power exponent. In addition, the power exponent of the head part of the distribution changes in time and Zipf's law holds only in a restricted period. We show that Zipf's law broke down due to both of Showa and Heisei great mergers and recovered due to population growth in middle-sized cities after the great Showa merger.

KEYWORDS: population, rank-size distribution, power-law distribution, lognormal distribution, Zipf's law, Gibrat's law, agent-based model

1. Introduction

Many empirical data which obey power-law distribution can often be observed in both natural and social sciences.^{1,2)} For example, the size distribution of lunar craters,³⁾ the relation between frequency and magnitude of earthquakes,⁴⁾ the size distribution of islands⁵⁾ and the cumulative probability distribution of personal income in Japan⁶⁾ are known to obey power-law distribution.

On the other hand, many researches have reported that empirical data obeying lognormal distribution are abundant around us. The lognormal distribution has the form

$$N(z) = \frac{1}{\sqrt{2\pi\sigma^2}z} \exp\left(-\frac{[\ln(z/T)]^2}{2\sigma^2}\right), \quad (1)$$

where σ and T are dispersion and average, respectively. For example, the fragmentation of glass rods,⁷⁾ income distribution of families and single individuals in U.S.,⁸⁾ the size distribution of barchan dunes,⁹⁾ food fragmentation by chewing¹⁰⁾ and the duration of disability for aged people¹¹⁾ are believed to obey Eq.(1) approximately.

The origin of those characteristic distributions is explained by the random multiplicative process which is often used to mimic the growth process of living organisms.¹²⁾ Let X_i be a physical quantity at time step i and suppose that its growth process is governed by the following relation: $X_i = \alpha_{i-1}X_{i-1}$. Here, α_i is the growth rate at the time step i . At the m -th step, X_m can be written as $X_m = X_0 \prod_{i=0}^{m-1} \alpha_i$, where X_0 is the initial quantity. Thus, when we assume that α_i is a random variable independent of X_i and m is sufficiently large, $\log X_i$ obeys the normal distribution due to the central limit theorem, which entails the lognormal distribution of X_i . This process is often called Gibrat's process or Gibrat's law, which is so common to many complex systems that one may say that the default distribution is the lognormal distribution. Indeed, if we introduce additional term, such as a noise term, into the Gibrat's process, X_i obeys power-law distributions.^{13,14)}

This paper focuses on the population distribution of cities in Japan. The population distribution within a given region sometimes shows power-law behavior. Auer-

bach firstly reported that the rank-size distribution for population of cities obeys a power-law distribution.¹⁵⁾ This means that when we order the cities by population and plot the rank $R(x)$ against its corresponding population x , the relation between $R(x)$ and x can be approximated by

$$\log R(x) = a - b \log x, \quad (2)$$

where a and b are fitting parameters. As for other municipalities, for example, Sasaki *et al.* recently reported that the rank-size distributions of towns and villages can be well approximated by lognormal distributions while that of cities obeys power-law distribution in Japan.¹⁶⁾

Zipf reported that the power exponent b is approximately 1 in the case of cities, so that the special case $b = 1$ is generally called Zipf's law.¹⁷⁾ Since many empirical data, such as the frequency of words in a literature and the income of companies,¹⁸⁾ obey Zipf's law, it is believed to be universal regularity.

However, we can easily find that Zipf's law in population distribution of cities is not universal.¹⁹⁾ For example, Figs. 1 (a) and (b) show the rank-size distribution of top 300 cities of U.S.A. in 2002 and that of top 267 cities of Brazil in 2006, respectively. Power exponents are $b = 1.338 \pm 0.004$ and $b = 1.230 \pm 0.005$, respectively, which are different from $b = 1$ predicted by Zipf's law, although the distributions obey power-law behavior. Here, we obtain those power exponents by the least square linear regression after plotting log of rank against log of population. In addition, we often find that the rank-size distribution does not exhibit power-law behavior.¹⁹⁾ Even if the rank-size distribution obeys power-law distribution, it is easily expected that power exponent b changes in time due to migration, a change of birth rate, and so on.

Some stochastic models have been proposed to explain Zipf's law. For example, Simon's model explains the emergence of Zipf's law in the rank-frequency distribution of words in literature.²⁰⁾ The point of this model is that, for adding a new word to a text, the probability that a word is repeated is proportional to the number of its previous occurrence. This model explains that the

*E-mail address: kuninaka@phys.chuo-u.ac.jp

rank-frequency distribution obeys power-law distribution with the exponent less than or equal to unity, which depends on the probability to choose the newly added word. On the other hand, Cancho has developed a model to explain the case that the exponent becomes larger than unity.²¹⁾ Recently, by modifying Simon's model, Zanette and Montemurro have developed a more realistic model to reproduce Zipf's law in the rank-frequency distribution of words, which explains the exponents in some cases of different languages.²²⁾

In this paper, we investigate the time evolution of the rank-size distribution of cities in Japan to show how the power exponent b changes after World War II. In addition, we show that Zipf's law holds only in a restricted period to explain why Zipf's law breaks down in Japan from our results of data analysis and simulation. Our data analysis is based on the census data from 1950 to 2006 which obtained from Statistics Bureau, Ministry of International Affairs and Communications, Japan,²³⁾ and data book from Japan Statistical Association.²⁴⁾

The organization of this paper is as follows. In the next section, we show our data analyses about the time evolution of the rank-size distribution for population of cities and the power exponent of its head part. Section 3 is devoted to modelling of population migration to explain the time evolution of the power exponent. In §4, we discuss our results of data analyses and simulation. The final section summarizes our results.

2. Data Analyses

Figure 2 shows the rank-size distributions for cities of Japan in 1950, 1960, 2000 and 2005, respectively. In each year, the rank-size distribution can be divided into two parts. For example, the distributions in 2000 and 2005 are clearly divided into two parts around 5.0×10^5 in population. Thus, the head and the tail part of each distribution can be fitted by discrete power-law distribution functions.

The slopes of head parts of the rank-size distributions change significantly from 1950 to 1960. This is mainly due to the fact that the number of cities drastically increased from 248 to 565 in the the great Showa merger from 1955 to 1960. From 2000 to 2005, the slope of head parts slightly increases from 1.027 ± 0.004 to 1.080 ± 0.004 , although the two distributions globally seem to be similar. Also in this case, the change of slopes of head parts is affected by the increase of the number of cities due to the great Heisei merger from 2000. Thus, the power exponent of the distribution changes easily by the great merger of municipalities.

Next, we investigate how the power exponent of head part of the rank-size distribution changed in time after World War II. Figure 3 shows the time evolution of the power exponent b from 1950 to 2006. Error bars which are almost invisible on data marks are standard deviation obtained by the least-squares linear regression. This figure shows that the power exponent b drastically changes during the two great mergers both in Showa and Heisei era. After the great Showa merger finished in 1960, the power exponent b shows monotonic decrease and approaches unity. The power exponent b keeps the value

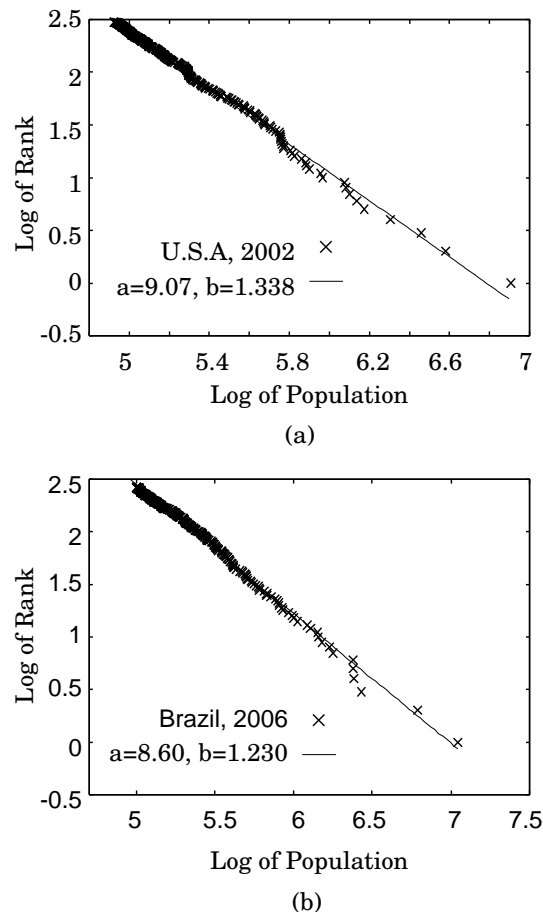


Fig. 1. Rank-size distributions for population of cities (a) in U.S.A, 2002 and (b) in Brazil, 2006.

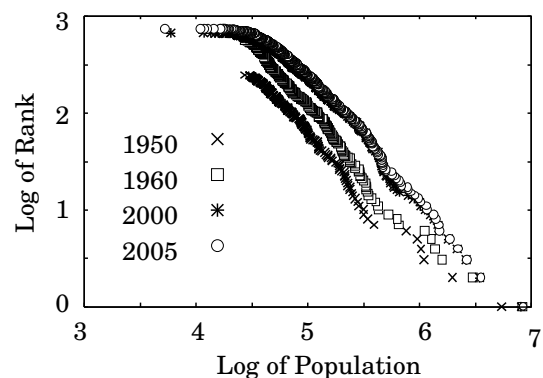
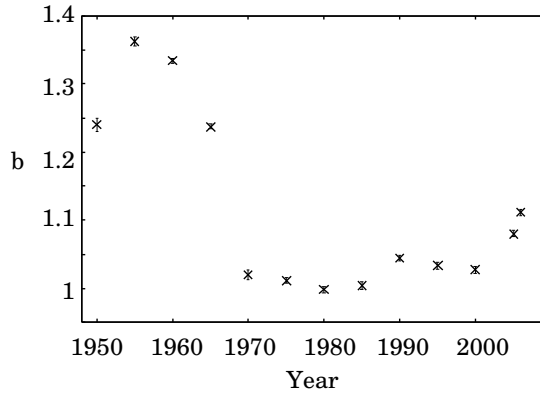


Fig. 2. Rank-size distribution of cities in Japan from 1950 to 2005.

near unity until the great Heisei merger starts in 2000. Thus, it is shown that Zipf's law holds only in the period from 1970 to 2000 in Japan.

Here, we should comment on the fitting range to obtain b . As we can see in Fig.2, the range of the head part is not so large at each year. Thus, all the values in Fig.3 are obtained by regression within about one order of magnitude.

To explain the relaxation of b to unity, we investigate the time evolution for the growth rate of population. To

Fig. 3. Time evolution of power exponent b from 1950 to 2006.

calculate the growth rate, at first, we categorize cities into some groups. The n -th group ($2 \leq n \leq 4$) is composed of 80 cities, the rank of which ranges from $80n - 139$ to $80n - 60$ at a given year, while the first group ($n = 1$) is composed of 20 cities, the rank of which ranges from 1 to 20. Note that the constituents of each group changes because the rank of cities usually changes at every census year.

We define the growth rate $P^n(t)$ of the n -th group at a census year t as

$$P^n(t) = \frac{\sum_{i=1}^{n_c} x_i^n(t)}{\sum_{i=1}^{n_c} x_i^n(t - \Delta t)}, \quad (3)$$

$$n_c = \begin{cases} 20 & (n = 1) \\ 80 & (n = 2, 3, 4), \end{cases}$$

where $x_i^n(t)$ is the population of the i -th city which belongs to the n -th group, and Δt is taken as $\Delta t = 5$ which is the interval between two successive census years in Japan. Figure 4 is the time evolution of the growth rate $P^n(t)$ of each group from 1960 to 2000. The growth rate of the first group shows global decrease, while those of other groups have apparent peaks in 1970 or 1975. This may be attributed to the following two factors: (i) the migration from the big cities to their satellite cities or the countryside, such as “U turn phenomena” or “I turn phenomena”, which is remarkable after 1970²⁵⁾ and (ii) population increase due to the second baby boom in the first half of the 1970s.

From these results, we can understand the time evolution of the power exponent b in Fig. 3 by the following scenario:

- (1) Before the great Showa merger starts, the power exponent b has the value near unity.
- (2) Due to the increase of the number of cities by the great Showa merger, the power exponent b increases.
- (3) After the merger, under the circumstance that the increase of the number of cities is not so large, the population of cities whose ranks range from 20 to 260 increases, which results in the decrease of b .
- (4) The power exponent b remains the value near unity until the great Heisei merger starts in 2000.

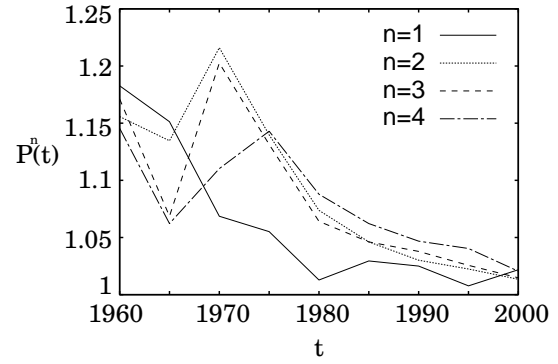


Fig. 4. Time evolution of growth rate from 1960 to 2000.

Table I. Number of cities of each year.

Year	Number of cities
1950	254
1960	561
1970	588
1980	647
1990	656

Here we would like to comment on why the period in which Zipf's law held continued for about 30 years. The head part of the rank-size distribution of cities consists of the groups with $n \geq 2$. From Fig. 4, we can easily find that the growth rates of these groups have almost the same value after 1975. In addition, the number of cities showed slow increase after 1960, while it had shown fast increase between 1950 and 1960²³⁾ (see Table. I). This may cause the stability of the power exponent after Zipf's law holds and prevent the exponent from taking the value less than unity.

3. Modelling on Population Migration

In this section, we construct a model for the population migration to reproduce the increase of b due to the merger of municipalities and its convergence to unity after the merger. Our model is based on an agent-based model which consists of 3500 sites corresponding to all the municipalities. Each site has a uniform random number between 0 and 1 as the initial population. The basic procedure of one simulation step is summarized as follows:

- (1) We randomly choose a source site m with the population N_m .
- (2) We choose a group of sites, $G_{N < N_m}$ or $G_{N > N_m}$, which are the groups of the sites whose population N are less and more than N_m , respectively. The probability to choose $G_{N < N_m}$ is α (migration parameter) while that to choose $G_{N > N_m}$ is $1 - \alpha$.
- (3) Among the group of sites chosen in the previous step, we randomly choose the destination site n for migration.
- (4) P_{mn} percent of N_m are transferred to the site n , so that the populations of sites m and n vary in quantity as $N_m - P_{mn}N_m$ and $N_n + P_{mn}N_m$, respectively.

In the second step, the migration parameter α is introduced to describe the tendency that people migrate to less populated area from large cities which was evident after the high economic growth from 1960 to early 1970s.²⁵⁾ In addition, P_{mn} is randomly chosen in the range from 0 to 20. We iterate this procedure 10^6 times in our simulation. Sample average is taken over 10 different initial population distributions for all the sites.

When the population of a given site becomes larger than 0.95, we regard the site as a city. Once a site is promoted to a city, the site will not be demoted to a smaller municipality such as towns and villages. This rule corresponds to a part of the Local Autonomy Law of Japan which says that municipalities must have a population of 50,000 or more to be promoted to cities.³⁰⁾ Our model does not distinguish between towns and cities. Thus, if a site does not belong to cities, we henceforth call the site as a “town”.

After the first migration of 10^6 simulation steps, we merge some municipalities according to the following procedure. At first, we randomly choose two sites to merge among all the sites. When both of them are not cities, we merge them to produce a new city if the sum of those populations becomes larger than 0.95, while we merge them to produce a town if the sum is less than 0.95. On the other hand, when at least one site is a city, we merge those two sites with the probability $\beta = 0.5$ to become a new city. The probability β is introduced due to the fact that the frequency of the merger of towns was much larger than that of cities. We iterate this merging process until the number of cities increases by 77 on average rather than that when the first migration stage is finished. In our model, the increase of the number of cities affects the power exponent after the merger. In general, the power exponent increases with the increase of the number of cities generated by the merger.

4. Simulation Results

At first, we investigate the convergence of the rank-size distribution of cities generated by our model. Figure 5 shows the rank-size distributions of cities at 10^5 , 10^6 , and 10^7 simulation steps, respectively. To obtain these results, the value of α is fixed at $\alpha = 0.3$. This figure shows that the rank-size distribution converges to the stationary power-law distribution with the power exponent $b = 1.012 \pm 0.002$. When the number of sites is more than 3500, our model needs longer simulation steps for the convergence to the power-law distribution with $b = 1$. Thus, our model can reproduce the power-law distribution of cities which converges to Zipf’s law. However, in our model, the number of cities keeps increasing after the power exponent becomes $b = 1$, which slightly increases the power exponent.

Secondly, we investigate how the great merger affects the rank-size distributions of cities through the time evolution of the power exponent b . In this simulation, we carry out the first population migration of 10^6 simulation steps. After that, we merge some of those sites, followed by the second population migration of 7×10^5 simulation steps. Figure 6 shows the time evolution of the rank-size distribution of cities. The dotted line shows the

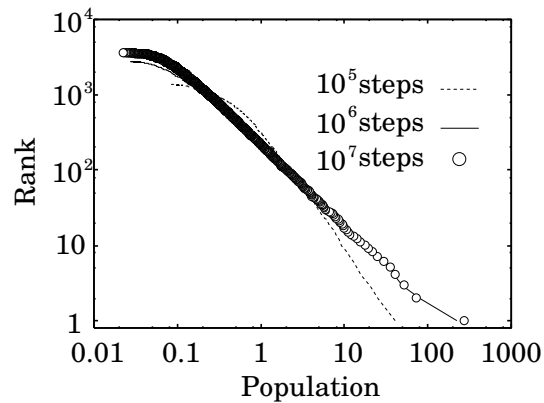


Fig. 5. Time evolution of rank-size distribution of cities without merging process.

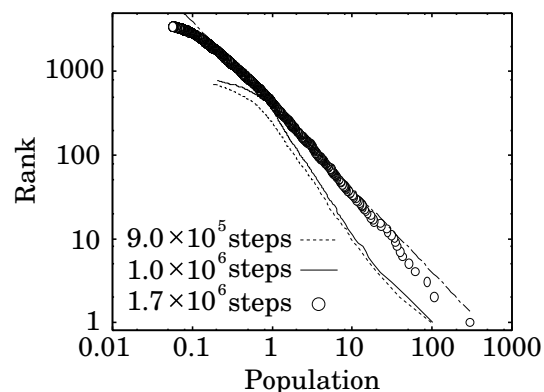


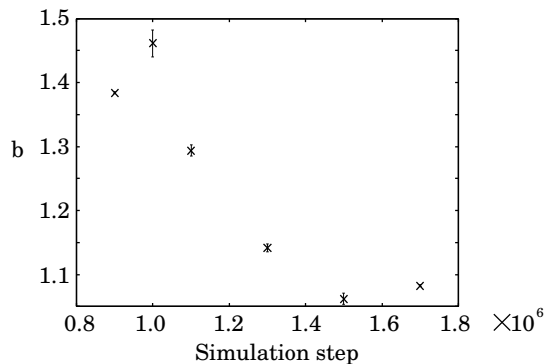
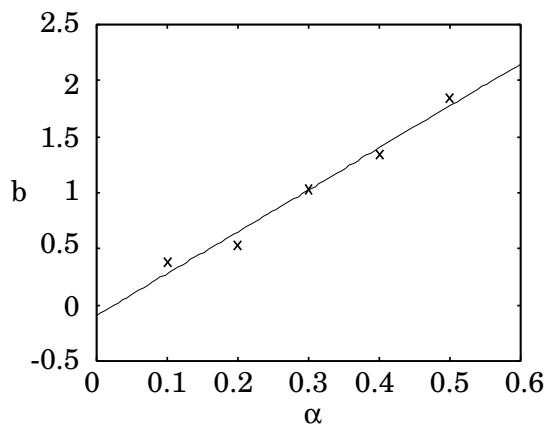
Fig. 6. Time evolution of rank-size distribution of cities with merging process.

distribution after the first migration stage was finished. The solid line shows the distribution after a merger of 200 sites. The open circles show the distribution after the second migration stage was finished, which can be fitted by the power-law distribution with the exponent $b = 1.081 \pm 0.001$ denoted by the dash-dotted line. Here we find that the distribution approaches the power-law distribution with the exponent $b = 1$ after the merger.

We show the relation between the power exponent b and the simulation step in Fig. 7. Error bars which are almost invisible on a few data marks are standard deviation obtained by the least-squares linear regression. Data point at 10^6 steps shows the power exponent b after the merger has finished. We find that b converges to unity after the increase of b due to the merger. Thus, our model can reproduce the time evolution of b qualitatively.

Finally, we investigate how α affects the final distribution. Figure 8 shows the relation between α and the power exponent b at 10^6 simulation steps. The solid line is the regression line: $b = 3.7\alpha - 0.09$. This result indicates that α determines the power exponent b of the final power-law distribution. For the convergence to Zipf’s law, this model requires $\alpha = 0.3$.

Here we would like to comment on the effect of initial population distribution on the final distribution. When we give $N_i = 1.0$ for initial value of all the sites, the power

Fig. 7. Time evolution of power exponent b .Fig. 8. Relation between α and b .

exponents of the resulting distributions do not show large difference.

5. Discussion

Let us discuss our results. From Fig.3, we find that Zipf's law holds for 25 years and breaks down due to the great Heisei merger. Naturally arises a question whether Zipf's law held also before 1950. However, during World War II, the number and distribution of people must have shown large fluctuation due to the great air campaigns against large cities such as Tokyo and Osaka, and evacuations from large cities to countrysides. Under a circumstance that the population distribution is unstable, it may be of little importance in discussing whether Zipf's law holds because there is a possibility that the distribution no more obeys power-law one.

We have found that the power exponent b approached unity after the great Showa merger had finished. Some theoretical explanations for the emergence of Zipf's law have been proposed in literature.^{20, 26, 27)} Among them, Gabaix showed that Gibrat's law in the population growth of each city is necessary for the emergence of Zipf's law.^{26, 28)} Here, Gibrat's law means that different cities grow randomly with the growth rate independent of the population of cities. We investigated the relation between the growth rate $P(t) \equiv x(t)/x(t - \Delta t)$ and the population $x(t)$ for all cities at $t = 1970$ (Fig.9). The

reason why we focus on $t = 1970$ is that the convergence to Zipf's law can be seen in this period (Fig.3). Here, we can run the regression,

$$\log P(1970) = 0.11 - (0.017 \pm 0.007) \log x(1970), \quad (4)$$

which has a slight slope, although it is supposed to become 0 if Gibrat's law holds. In addition, the dispersion of growth rate becomes rather large around $\log x(1970) = 4.5$, so that we cannot clearly see whether Gibrat's law holds or not. In the case of all municipalities, Sasaki *et al.* reported that the slope becomes almost 0 from 2000 to 2005, although it has a slight slope.¹⁶⁾

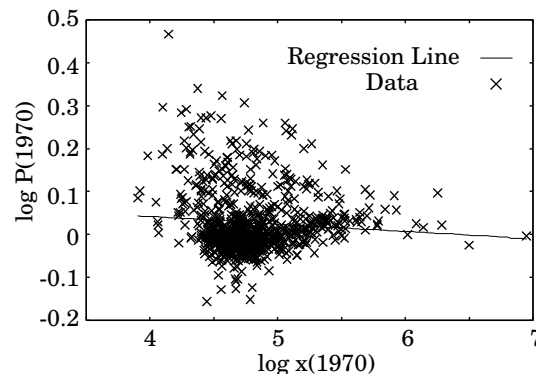


Fig. 9. Relation between log of growth rate and log of population for cities from 1965 to 1970. Solid line is Eq.(4)

As we referred in §1, a random multiplicative process with Gibrat's law generates lognormal distribution. The rank-size distribution for population in all municipalities shows the double-Pareto distribution, which consists of lognormal body with power-law tail.^{14, 16, 29)} Thus, it is no wonder that the regression line for the relation between the growth rate and the population has a non-zero slope. Because the rank-size distribution of cities is the tail part of that of municipalities, it may have a non-zero slope. Thus, in the case of population of cities, Gibrat's law may be just necessary condition for the emergence of Zipf's law.

To obtain the power exponent b of rank-size distributions for cities, we adopt the least-squares linear regression to fit those distribution functions by Eq.(2). Although this method is used frequently in literature, it is known that the method has several problems.²⁾ To obtain more reliable estimate for b , other estimation methods such as the maximum likelihood method²⁾ may be better.

In §3, we have constructed the agent-based model to explain the emergence and the breakdown of Zipf's law in the rank-size distribution of cities. This model can reproduce Zipf's law which is observed in the process that many entities exchange physical quantities among them. We often find Zipf's law in some phenomena without an apparent exchange process such as word frequencies in literature and the relation between the frequency and the magnitude of earthquakes. However, even for these cases there may be some hidden exchange processes such as

words in/out of fashion and the accumulation/relaxation of the crust stress due to the plate tectonic movement. Moreover, exchange processes are almost universal in the economic world as well as social world. Hence we believe that the present model can be applicable to other problems as well.

In §4, we carried out a simulation of population migration to explain the time evolution of the power exponent b in Fig. 3. In Fig. 6, after the merger of municipalities, the rank-size distribution shifts towards upper direction in all the region. If we use the value of β smaller than 0.5, the distribution shifts towards upper direction in the region whose population is less than about 1.8. Consequently, small value of β causes a decrease of the range in which the distribution can be fitted by a single power-law distribution.

The rank-size distribution of all municipalities has a lognormal body and a power-law tail,¹⁶⁾ which is observed also in our simulation. This type of distribution can be observed in the agent-based simulation of exchanging quantities on the small-world network,³¹⁾ which implies the possibility that the population migration network may have a small-world structure. To clarify the relevance, we need to analyse the population migration network between municipalities in detail.

6. Concluding Remarks

In conclusion, we have investigated the time evolution of the rank-size distribution for population of cities to show how the power exponent changes in time. The rank-size distribution shows that power-law behavior and the time evolution of the power exponent drastically changes when the great merger of municipalities occurs. After the great Showa merger finished, the power exponent converged to unity, which means that Zipf's law holds. We have explained the change of the power exponent by the growth rates of the categorized groups of cities in the point of view of migration.

We would like to thank M. Katori, T. Nakano, S. Tomita, N. Kobayashi, Y. Sasaki and T. Miyazaki for useful discussions. We would also like to thank Y. Aruka, A. Namatame and H. Hayakawa for their useful comments. Numerical computation was partially carried out at the Yukawa Institute Computer Facility. A part of this work is supported by a Grant-in-Aid for Young Scientists (B) of the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan (Grant

No.20740226).

- 1) M. E. J. Newman: *Contemp. Phys.* **46** (2005) 323.
- 2) A. Clauset, C. R. Shalizi, and M. E. J. Newman: *physics/0706.1062*.
- 3) R. B. Baldwin: *Astron. J.* **69** (1964) 377.
- 4) See, e.g., D. L. Turcotte: *Fractals and Chaos in Geology and Geophysics* (Cambridge University Press, Cambridge, 1992).
- 5) Y. Sasaki, N. Kobayashi, S. Ouchi and M. Matsushita: *J. Phys. Soc. Jpn.* **75** (2006) 074804.
- 6) H. Aoyama, W. Souma, and Y. Fujiwara: *Physica A* **324** (2003) 352.
- 7) T. Ishii and M. Matsushita: *J. Phys. Soc. Jpn.* **61** (1992) 3474.
- 8) E. W. Montroll and M. F. Shlesinger: *Proc. Natl. Acad. Sci. U.S.A.* **79** (1982) 3380.
- 9) O. Durán: Dr. Thesis, Fakultät Mathematik und Physik, Universität Stuttgart, Stuttgart (2006).
- 10) N. Kobayashi, K. Kohyama, Y. Sasaki and M. Matsushita: *J. Phys. Soc. Jpn.* **75** (2006) 083001.
- 11) O. Moriyama, H. Itoh, S. Matsushita and M. Matsushita: *J. Phys. Soc. Jpn.* **72** (2003) 2409.
- 12) E. W. Crow and K. Shimizu: *Lognormal Distributions: theory and applications* (Marcel Dekker, NY, 1988) p.4.
- 13) H. Takayasu, A. Sato, and M. Takayasu: *Phys. Rev. Lett.* **79** (1997) 966.
- 14) S. Tomita and Y. Hayashi: *Physica A.* **387** (2008) 1345.
- 15) F. Auerbach: *Petermanns Geogr. Mitt.* **LIX** (1913) 74.
- 16) Y. Sasaki, H. Kuninaka, N. Kobayashi and M. Matsushita: *J. Phys. Soc. Jpn.* **76** (2007) 074801.
- 17) G. Zipf: *Human Behavior and the Principle of Least Effort*. (Addison-Wesley, Cambridge, 1949).
- 18) We can find many reports about empirical data obeying Zipf's law at <http://www.nslj-genetics.org/wli/zipf/>.
- 19) K. T. Soo: *Reg. Sci. Urban. Econ.* **35** (2005) 239.
- 20) H. A. Simon: *Biometrika* **44** (1955) 425.
- 21) R. Ferrer i Cancho: *Eur. Phys. J. B* **47** (2005) 449.
- 22) D. H. Zanette and M. A. Montemurro: *J. Quant. Linguistics.* **12** (2005) 29.
- 23) <http://www.stat.go.jp/>.
- 24) Shikuchoson Jinko no Chouki Keiretsu(Long-term data for population of cities, wards, towns and villages) (Nihon Tokei Kyokai, Tokyo, 2005) [in Japanese].
- 25) Y. Arai, T. Kawaguchi and T. Inoue: *Nihon no Jinkou Idou(migration in Japan)* (Kokon Shoin, Tokyo, 2002) p.5 [in Japanese].
- 26) X. Gabaix: *Quart. J. Econ.* **114** (1999) 739.
- 27) K. Kawamura and N. Hatano: *J. Phys. Soc. Jpn.* **71** (2002) 1211.
- 28) Y. M. Ioannides and H. G. Overman: *Reg. Sci. Urban. Econ.* **33** (2003) 127.
- 29) M. Mitzenmacher: *Internet Math.* **1** (2003) 305.
- 30) <http://nippon.zaidan.info/seikabutsu/1999/00168/mokuji.htm>.
- 31) W. Souma, Y. Fujiwara and H. Aoyama: *cond-mat/0108482v1*.